© 2005 Schattauer GmbH

Meta-Analysis: a Unifying Meta-Likelihood Approach Framing Unobserved Heterogeneity, Study Covariates, Publication Bias, and Study Quality

D. Böhning
Division for International Health, Institute for Social Medicine, Epidemiology, and Health Economics, Charité Medical School Berlin, Free University Berlin/Humboldt University, Berlin, Germany

Summary

Objectives: This contribution provides a unifying concept for meta-analysis integrating the handling of unobserved heterogeneity, study covariates, publication bias and study quality. It is important to consider these issues *simultaneously* to avoid the occurrence of artifacts, and a method for doing so is suggested here.

Methods: The approach is based upon the meta-likelihood in combination with a *general linear nonparametric mixed model*, which lays the ground for all inferential conclusions suggested here.

Results: The concept is illustrated at hand of a meta-analysis investigating the relationship of hormone replacement therapy and breast cancer. The phenomenon of interest has been investigated in many studies for a considerable time and different results were reported. In 1992 a meta-analysis by Sillero-Arenas et al. [1] concluded a small, but significant overall effect of 1.06 on the relative risk scale. Using the meta-likelihood approach it is demonstrated here that this meta-analysis is due to considerable unobserved heterogeneity. Furthermore, it is shown that new methods are available to model this heterogeneity successfully. It is argued further to include available study covariates to explain this heterogeneity in the meta-analysis at hand.

Conclusions: The topic of HRT and breast cancer has again very recently become an issue of public debate, when results of a large trial investigating the health effects of hormone replacement therapy were published indicating an increased risk for breast cancer (risk ratio of 1.26). Using an adequate regression model in the previously published meta-analysis an adjusted estimate of effect of 1.14 can be given which is considerably higher than the one published in the meta-analysis of Sillero-Arenas et al. [1]. In summary, it is hoped that the method suggested here contributes further to a good meta-analytic practice in public health and clinical disciplines.

Keywords

Evidence-based medicine, C.A.MAN, meta-analysis, meta-regression, meta-likelihood, nonparametric maximum meta-likelihood, publication bias, unobserved heterogeneity

Methods Inf Med 2005; 44: ■–■

1. Introduction and Background

In all empirical sciences there exists an enormous body of empirical knowledge for a given question of interest. This knowledge has often been collected in numerous studies and empirical investigations by means of experimental studies, clinical trials or observational studies. Typically, these individual findings are buried in the scientific literature, in registries, or some other form of documentary source. From here, they must be retrieved and relevant information extracted, and finally, statistically analyzed by an appropriate methodology. We have entered the territory of *meta-analysis*. Numerous publications in the area underline that meta-analysis has become a central role in the collection, analysis and evaluation of findings in any empirical science. Before we proceed to develop a general statistical framework for coping with controversial issues in a more universal framework, we hope to initiate interest in the question by recalling a recent debate on the public health issue of hormone replacement therapy. Hormone replacement therapy is applied to achieve positive effects in many respects for women near and after menopause. It has been questioned, however, for a longer period if it relates to the occurrence of breast cancer.

In a recent evening news broadcast *Die Tagesthemen*[a] the following event was reported which translates as follows: **"New Doubts on Hormone Replacement Therapy.** Due to arising care for the participating women a hormone study *(the WHI-study)* in the USA including 16,000 older women was terminated earlier than scheduled. The objective of the study was to investigate the benefit of Oestrogenes and Progestines for females in or after menopause. As it turned out that the risks occurring due to the daily dosages of hormone are larger than their potential benefits, the US health administration NIH declared termination of the study three years prior to the end of the designed study period. Specifically, more cases of breast cancer, myocardial infarction, stroke, and blood clothing in the lung had occurred in the exposed group (under hormone treatment) than in the control group (no exposure) which was given a placebo presumingly of no effect." (Die Tagesthemen, ARD 10:30 p.m.). Details of the before mentioned study can be found elsewhere [2]. For breast cancer the study provided a risk ratio of 1.26. This seems to be a minor effect. However, one should keep in mind that even a small elevated relative risk can have large public health effects if the size of the exposed population is large. To demonstrate let us assume that in a population with a baseline risk of breast cancer of 1 in 100 there are 1,000,000 under hormone replacement therapy (HRT) leading to 2,600 additional cases due to HRT (using the risk ratio of the above mentioned study). Even with a diminished assumption of 1 in 1000 (the baseline risk e.g. the risk in the placebo arm of the WHI-study was, with 124 breast cancer cases in 8,102 women under risk, far above this number [2]) one can expect about 260 additional cases due to HRT. Thus, even small effects are of considerable importance

[a] This evening news belongs to Channel One of public German television and has typically leading participant quota.

Received: July 31, 2003; accepted: April 21, 2004

Author's personal copy for non-commercial use only

998 BERLIN DER TAGESSPIEGEL / SEITE 13

Medikamenten-Studie an Frauen: Gefährdet ein Arzt seine Patienten?

Hormontherapie kann Brustkrebs auslösen, dennoch will FU-Professor weiterforschen

VON INGO BACH

Straffere Haut und schönes Haar dank Hormonbehandlung versprach Professor Dieter Felsenberg Ende des vergangenen Jahres 20 000 Berliner Frauen. Er hatte sie angeschrieben, die Adressen erhielt er vom Landeseinwohneramt – eine legale Amtshilfe, wie er versichert. Mit seinem Brief über die Segnungen der Hormontherapie suchte Felsenberg, Radiologe am Universitätsklinikum Benjamin Franklin in Steglitz, nach Probandinnen für eine Studie zur Wirkung eines neuen Hormonpräparats auf Haut und Haar.

Nun aber sieht sich Felsenberg mit harscher Kritik konfrontiert: Er gefährde die Frauen, heißt es. Hintergrund der Vorwürfe ist eine Langzeituntersuchung in den USA, bei der 16 000 Frauen im Alter von 50 bis 79 Jahren mit einer ähnlichen Wirkstoffkombination behandelt wurden, wie sie jetzt auch Felsenberg testen will. Diese US-Studie wurde im Sommer 2002 vorzeitig abgebrochen, weil der Schaden den möglichen Nutzen deutlich überwog. Die Probandinnen erkrankten häufiger an Brustkrebs, erlitten Herzinfarkte und Lungenembolien.

Auch das Bundesinstitut für Arzneimittel und Medizinprodukte in Bonn sieht im Lichte dieser Resultate eine Hormontherapie, wie sie bisher auch gesunden Frauen in den Wechseljahren verschrieben wurde, kritischer: Diese Patientinnen sollten ihren Arzt fragen, „ob und in welcher Form die Fortsetzung einer Hormonersatztherapie noch weiterhin sinnvoll" sei.

Die Ergebnisse der US-Kollegen lassen auch Dieter Felsenberg nicht unbeeindruckt: „Vor diesem Hintergrund hätte ich es mir noch einmal sehr genau überlegt, ob ich diese Untersuchung machen soll", sagt er. Doch das positive Votum der Ethikkommission des Uniklinikums, die jede Pharmastudie begutachtet (siehe Kasten), hat er schon vor Veröffentlichung der amerikanischen

DER TAGESSPIEGEL
10. Januar 2003

Studie erhalten. Er setzt deshalb seine Forschung fort. „Ich möchte klären, ob der Präparat-Hersteller damit werben kann, dass die Frauen dadurch eine glattere Haut bekommen oder schöneres Haar."

150 Frauen sollen dafür ein Jahr lang das Medikament testen. Er habe die Teilnehmerinnen über die Resultate aus den USA unterrichtet und sie vor die Wahl gestellt, den Versuch abzubrechen. Keine sei zurückgetreten, sagt Felsenberg.

Im Gegensatz zu den amerikanischen Ärzten hält der Professor aus Berlin das Risiko für vertretbar und den Nutzen für groß.

Allein durch die Voruntersuchungen habe er drei Frauen entdeckt, die einen Brustkrebs im Frühstadium entwickelt hatten und nun mit einer hohen Heilungschance behandelt werden könnten. Die Wahrscheinlichkeit, durch das Medikament an Brustkrebs zu erkranken, liege im Promille-Bereich. Laut der US-Studie stieg der Anteil von Brustkrebs-Erkrankungen unter 10 000 Frauen von 30 auf 38 Fälle an.

Wolfgang Becker-Brüser, Geschäftsführer des „Arznei-Telegramms", das den Ruf als seriöses Fachblatt genießt, widerspricht: „Bei dem theoretisch möglichen Markt von vier Millionen Frauen, die in der Bundesrepublik Hormonpräparate erhalten, wären das immerhin rund 3200 durch das Medikament ausgelöste Brustkrebsfälle." Und für eine durch die Studie erkrankte Patientin sei die statistische Wahrscheinlichkeit belanglos. Für sie zähle nur, dass sie Krebs hat.

ETHIKKOMMISSION UND PATIENTENSCHUTZ

Jede Pharmastudie muss von einer Ethikkommission begutachtet werden. Allerdings ist ihr Votum nicht bindend. Der Arzt, der die Studie leitet, hat immer das letzte Wort und kann seine Versuchsreihe auch gegen das Urteil der Kommission durchführen. Das aber wird er im Normalfall nicht tun, sagt Joachim W. Dudenhausen, Dekan der Charité. Nähmen die Patienten im Laufe einer solchen Studie Schaden, hätte der Arzt im Klagefall keine Unterstützung und müsste auch um seine Reputation fürchten. In der Kommission sitzen Ärzte, Theologen, Juristen und Patientenvertreter. Sie wägen den möglichen Schaden – etwa durch Nebenwirkungen – gegen den Nutzen ab. Manchmal zieht das Gremium auch externe Fachmediziner hinzu. Außerdem prüft die Kommission, ob die Teilnehmer verständlich über Risiken aufgeklärt wurden und ob der Auftraggeber oder Leitende Arzt eine Haftpflichtversicherung abgeschlossen haben.

Trotz möglicher gesundheitlicher Risiken für die Teilnehmer könne auf Studien nicht verzichtet werden, sagt Dudenhausen. „Ohne sie ist ein medizinischer Fortschritt nicht möglich." I.B.

Fig. 1 Critical report on a clinical trial using HRT in the daily DER TAGESSPIEGEL, January 27, 2003. Title translates as: Medication Study of Women: Does a Medical Doctor Put his Patients at Risk? Subtitle translates as: Hormone Therapy Can Trigger Breast Cancer, nevertheless, a Professor at the Free University Berlin Will Continue his Research

when the exposure is widespread in the community.

This finding of the WHI-study has not only been taken up by the media as important health news, even medical investigators executing clinical trials with a hormone replacement therapy arm were taken by surprise by the results of the WHI-trial (see Fig. 1). In this case, a medical study was publicly criticized for using HRT, though it would have been known that an excess risk for developing breast cancer exists. Interestingly, in its response the medical team pointed out that they would have *not* executed the trial in this fashion if the excess risks had been known prior to the time of the beginning of their study. In addition, the medical team pointed the attention of the reader to the agreement of the hospital's ethical committee. However, as we will argue in this contribution, appropriate analysis of the body of evidence would have flagged sources of excess risks a considerable time earlier.

Indeed, if we reconsider the meta-analysis by Sillero-Arenas et al. [1] and use the appropriate tools in a secondary analysis, we find an estimate of 1.137 on the relative risk scale. This finding occurs since the meta-analysis at hand experiences considerable *unobserved heterogeneity,* which has previously been ignored (leading to a diminished effect estimate of 1.06). We argue further in this contribution that this form of heterogeneity can be successfully linked to covariates observed in the study base provided by SA, namely the study type (cohort or case-control) and whether in the study the estimate of effect has been adjusted for potential confounders. It can be furthermore established that these covariates correlate with the size of the effect measure. Consequently, an odds ratio adjusted for study type and confounder treatment seems to be more appropriate, leading to the one given above.

In this contribution, we outline a general concept based on what is called *the meta-likelihood,* which provides a unifying approach to deal with several typical problems in the area of meta-analysis: *study covariates* and *unobserved heterogeneity, publication bias* and *study quality.*

2. An Application: The Data of the Meta-Analysis of Hormone Replacement Therapy and the Occurrence of Breast Cancer

We would like to come back to a meta-analysis provided by Sillero-Arenas et al. [1] – throughout this paper abbreviated as SA – and point out that the finding of the recent trial [2] is not surprising and essentially agrees with the result of the meta-analysis. The meta-analysis [1] contained 23 case-control studies and 13 cohort studies. There was also one clinical trial mentioned that is not further considered here since relevant information could not be retrieved for this study.

2.1 Effect Measure and 95% Confidence Interval

Since there is a mixture of study designs the odds ratio appears to be the appropriate effect measure. The odds ratio is validly estimable in case-control *and* cohort studies, whereas the relative risk can only be validly estimated in cohort studies, though the dif-

For Review Purpose Only

ferences between the two are small when the baseline-risk is small. Therefore, odds ratios might be interpretable as relative risks. Odds ratios were available in the original meta-analysis and were provided with 95% confidence intervals (see Table 1).

2.2 Standard Error

In the original meta-analysis the standard errors associated with the log-odds ratios were not available. However, they can easily be reconstructed since the 95% C.I. is constructed as log(OR) ± 1.96 SE, where SE is the standard error for the log(OR). Let the two interval ends be denoted by U and L, then the standard error is found as $SE = (U - L)/(1.96 \times 2)$. This formula was used to construct the data provided in column 3 of Table 2.

2.3 Sample Size

Sample sizes were also provided in the original meta-analysis, but on different scales (number of persons for case-control studies and number of person-years for cohort studies). Since the sample size is related inversely to the standard error we will use 1/SE as a substitute for the sample size wherever this is needed.

2.4 Date of Data Collection

For most studies a time for the data collection is provided as well. If a time interval is given, we have used the mid-point.

2.5 Study Type

Two study types are used: cohort (13 studies) and case-control (23 studies). This information will be utilized and considered as a potential source of bias.

Table 1 Extracted data from meta-analysis by Sillero-Arenas et al. [1]

Study♣	OR	95% CI		Study-Type♥	Date of Data Coll.¥	Adjusted for Co-variates♠
1	1.11	0.38	1.19	1	72.0	1
2	0.97	0.49	1.92	1	68.0	1
3	2.15	0.71	6.49	1	71.0	1
4	0.82	0.60	1.20	1	71.5	2
5	0.90	0.66	1.22	1	72.0	0
6	0.89	0.60	1.32	1	73.0	0
7	1.10	0.80	1.90	1	74.0	*
8	1.30	1.00	1.70	1	72.0	*
9	0.77	0.58	1.02	1	78.0	*
10	1.58	1.09	2.28	1	77.5	1
11	0.55	0.32	0.94	1	76.0	2
12	0.70	0.30	1.60	1	69.5	1
13	0.90	0.50	1.70	1	77.5	0
14	0.73	0.58	0.90	1	78.5	2
15	0.90	0.50	1.30	1	81.5	2
16	1.03	0.90	1.20	1	76.5	1
17	1.84	1.27	2.68	1	84.0	2
18	0.74	0.51	1.08	1	77.5	1
19	1.02	0.75	1.38	1	77.5	1
20	1.00	0.90	1.20	1	81.0	2
21	0.96	0.75	1.22	1	82.5	2
22	1.03	0.62	1.69	1	83.0	1
23	1.20	0.98	1.47	1	83.5	1
24	1.30	1.00	1.70	0	55.5	1
25	1.38	0.81	2.33	0	*	1
26	1.97	0.50	1.78	0	*	1
27	0.38	0.16	0.88	0	65.0	0
28	2.50	1.60	4.00	0	77.0	*
29	1.38	0.72	2.65	0	59.0	1
30	0.32	0.18	0.57	0	78.0	0
31	0.62	0.33	1.14	0	70.0	0
32	1.18	1.04	1.35	0	81.0	*
33	1.59	1.18	2.10	0	80.0	1
34	0.59	0.38	0.83	0	59.0	1
35	1.11	0.99	1.24	0	80.5	1
36	1.74	1.10	2.74	0	79.0	2

♣ Studies are listed as in Table 1 by Sillero-Arenas et al. (1992).
♥ 1 corresponds to Case-Control-Study, 0 corrresponds to Cohort Study.
¥ A * indicates a missing value. ♠ Indicates how many covariates were adjusted for in each study.

For Int

Table 2 Log-odds ratios with associated standard errors from meta-analysis by Sillero-Arenas et al. [1]

Study♣	log OR	Standard Error	Study-Type♥	Date of Data Coll.¥	Adjusted for Co-variates♠
1	1.11	0.29	1	72.0	1
2	0.97	0.35	1	68.0	1
3	2.15	0.56	1	71.0	1
4	0.82	0.18	1	71.5	2
5	0.90	0.16	1	72.0	0
6	0.89	0.20	1	73.0	0
7	1.10	0.22	1	74.0	*
8	1.30	0.14	1	72.0	*
9	0.77	0.14	1	78.0	*
10	1.58	0.19	1	77.5	1
11	0.55	0.27	1	76.0	2
12	0.70	0.43	1	69.5	1
13	0.90	0.31	1	77.5	0
14	0.73	0.11	1	78.5	2
15	0.90	0.24	1	81.5	2
16	1.03	0.07	1	76.5	1
17	1.84	0.19	1	84.0	2
18	0.74	0.19	1	77.5	1
19	1.02	0.16	1	77.5	1
20	1.00	0.07	1	81.0	2
21	0.96	0.12	1	82.5	2
22	1.03	0.26	1	83.0	1
23	1.20	0.10	1	83.5	1
24	1.30	0.14	0	55.5	1
25	1.38	0.27	0	*	1
26	1.97	0.32	0	*	1
27	0.38	0.43	0	65.0	0
28	2.50	0.23	0	77.0	*
29	1.38	0.33	0	59.0	1
30	0.32	0.29	0	78.0	0
31	0.62	0.32	0	70.0	0
32	1.18	0.07	0	81.0	*
33	1.59	0.15	0	80.0	1
34	0.59	0.20	0	59.0	1
35	1.11	0.06	0	80.5	1
36	1.74	0.23	0	79.0	2

♣ Studies are listed as in Table 1 in [1].
♥ 1 corresponds to Case-Control.Study, 0 corresponds to Cohort Study
¥ A * indicates a missing value.
♠ Indicates how many covariates were adjusted for in each study.

2.6 Number of Covariates Adjusted for

When doing observational studies it is important to control for potential confounding covariates like age, BMI, etc. This information was available in the original meta-analysis and will also be considered as a potential source of bias. All these covariates are provided in Table 1.

3. Statistical Methods – An Approach Based upon the Meta-Likelihood

3.1 The Meta-Likelihood

Meta-analysis has become a standard tool in medical research. Recently, a number of excellent books have appeared [3-6] updating a number of earlier contributions [7-9]. In addition, special texts have appeared dealing with Bayesian approaches [10] or heterogeneity modeling [11]. In the following, a likelihood approach is used which appears to be widely accepted.

It is *assumed* that the effect measure $\hat{\lambda}_i$ for the i-th study (in the application it is the log-odds ratio) follows (at least approximately) a *normal distribution* with density of $\hat{\lambda}_i$:

$$\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{ - \tfrac{1}{2} (\hat{\lambda}_i-\lambda_i)^2/\sigma_i^2\} \qquad (1)$$

where λ_i is the *unknown* effect measure in study i and σ_i^2 is the *known* study variance (see column 3 of Table 2). Having k independent studies available this leads to the *meta-likelihood*

$$\prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{ - \tfrac{1}{2} (\hat{\lambda}_i-\lambda_i)^2/\sigma_i^2\} \qquad (2)$$

which will be the basis for all inferential conclusions. Now different models for the effect measure λ_i can be considered.

Use only

3.2 Homogeneity

This is the simplest model and also the most widely used. It assumes that all $\lambda_1 = \lambda_2 = ... = \lambda_k = \lambda$ coincide and maximizing the meta-likelihood (2) leads to the weighted mean $\hat{\lambda}_+ = w_1\hat{\lambda}_i + ... + w_k\hat{\lambda}_k / (w_1 + ... + w_k)$ of the *observed* effect measures of the k studies with $w_i = 1/\sigma_i^2$. This is also called the *pooled* or *fixed* effect estimate. It is particularly attractive since it combines all study estimates into *one* single measure (for details, see [4]). In addition, the variance of this estimate is readily available as $1/(w_1 + ... + w_k)$. Though attractive and simple, this approach is rarely appropriate in practice, since the assumption of homogeneity is often violated, and non-homogeneity or *heterogeneity* frequently occurs .

3.3 (Unobserved) Heterogeneity and the Nonparametric Meta-Likelihood

Effect-heterogeneity implies that a certain value for the effect is valid for some studies whereas for others a different value is correct. To demonstrate, it might be that the heterogeneity consists in two subpopulations, where one corresponds to a moderately harmful, the other to a more harmful effect, or, heterogeneity might consist out of three subpopulations, one corresponding to a harmful, the other to a beneficial, and the third to a null-effect. The latter example is particularly misleading when a simple, weighted mean, which might take on a value near the null-effect, is computed. How can such a situation be validly captured by means of a model? Typically, recent approaches concentrate on *random effects* models. These can be best illustrated as follows. One supposes that the studies are sampled from a population with a non-homogeneous effect pattern, in other words, there are a number of components experiencing different sizes in the effect. One can think of a distribution *P* according to which sampling of studies takes place. It is no limitation to assume that this distribution is discrete giving mass p_j to effect size λ_j, where j corresponds to the component in the population, j = 1, ..., m, where m is the (unknown) number of components. It is assumed that the membership of each study to the associated subpopulation is unknown. The pair $(\hat{\lambda}_i, \mathbf{z}_i)$ contains the observed effect measure $\hat{\lambda}_i$ of the i-th study and the *unobserved* indicator vector $\mathbf{z}_i$ with exactly one 1 in the j-th position, say, indicating that the i-th study belongs to the j-th subpopulation. The corresponding *unobserved* meta-likelihood is

$$\prod_{i=1}^{k}\prod_{j=1}^{m} p_j^{z_{ij}} \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\tfrac{1}{2}(\hat{\lambda}_i-\lambda_j)^2/\sigma_i^2\}\right]^{z_{ij}} \quad (3)$$

for which *closed form* solutions for p_j and λ_j exist. Unfortunately, z_{ij} are not known, so that the marginal likelihood (margin over the latent variable) is appropriate to be used:

$$\prod_{i=1}^{k}\sum_{i=1}^{m} p_j \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\tfrac{1}{2}(\hat{\lambda}_i-\lambda_j)^2/\sigma_i^2\} \quad (4)$$

This meta-likelihood is called *observed* and needs to be maximized in the parameters $\lambda_1, ..., \lambda_m, p_1, ..., p_m$. Note there are now $2m - 1$ parameters since the non-negative weights $p_1,...,p_m$ are summing up to 1. Note that if m = 1 this meta-likelihood reduces to the one given under homogeneity. The *unobserved* meta-likelihood (3) and the *observed* meta-likelihood (4) are connected by means of a many-to-one mapping which maps the pair $(\hat{\lambda}_i, \mathbf{z}_i)$ onto $\hat{\lambda}_i$. We have that unobserved and observed meta-likelihood are connected via

$$\prod_{i=1}^{k}\sum_{\mathbf{z}_i}\prod_{j=1}^{m} p_j^{z_{ij}} \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\tfrac{1}{2}(\hat{\lambda}_i-\lambda_j)^2/\sigma_i^2\}\right]^{z_{ij}}$$
$$= \prod_{i=1}^{k}\sum_{j=1}^{m} p_j \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\tfrac{1}{2}(\hat{\lambda}_i-\lambda_j)^2/\sigma_i^2\} \quad (5)$$

where the first summation is taken over all possible vectors $\mathbf{z}_i$ (for each i there are exactly m of those) having a single 1 at one position. The result (5) is one of the milestones of the EM-algorithmic theory [12]. Estimation of the $2m - 1$ parameters can be readily accomplished with the EM algorithm. The latter uses expected values $e_{ij} = E(Z_{ij}|\hat{\lambda}_i, \lambda_j, p_j)$ in the *unobserved* meta-likelihood (3) and maximization of this *expected, unobserved* meta-likelihood provides *new estimates* for p_j and λ_j:

$$\hat{p}_j^{new} = \frac{1}{k}\sum_{i=1}^{k} e_{ij} \quad \textit{and}$$

$$\hat{\lambda}_j^{new} = \sum_{i=1}^{k} w_i e_{ij} \hat{\lambda}_i \,/\, \left[\sum_{i=1}^{k} w_i e_{ij}\right] \quad (6)$$

This is the *M-step*. It remains to provide the *conditional* expected values $e_{ij} = E(Z_{ij}|\hat{\lambda}_i, \lambda_j, p_j)$. Let Λ_i be the random variable with realization $\hat{\lambda}_i$. Then

$$E(Z_{ij} \mid \Lambda_i = \hat{\lambda}_i, \lambda_j, p_j) = Pr(Z_{ij}=1 \mid \Lambda_i = \hat{\lambda}_i, \lambda_j, p_j) = Pr(\Lambda_i = \hat{\lambda}_i \mid Z_{ij}=1; \lambda_j, p_j) / \sum_{j'=1}^{m} Pr(\Lambda_i = \hat{\lambda}_i \mid Z_{ij'}=1; \lambda_{j'}, p_{j'})$$

where Bayes theorem was used in the last equation. Therefore, we have

$$e_{ij} = \varphi(\hat{\lambda}_i - \lambda_j)/\sigma_i^2)\, p_j \,/\, \sum_{j'=1}^{m} \varphi(\hat{\lambda}_i - \lambda_{j'})/\sigma_i^2)\, p_{j'} \quad (7)$$

where φ is the standard normal density. This completes the *E-step*.

The EM-algorithm proceeds by cycling between steps (6) and (7). C.A.MAN, a software tool freely available from the author's homepage can be used for computational practice. Details on the nonparametric mixture likelihood approach can be found in [11]. Note that this approach models the background heterogeneity more completely than other approaches like the one by DerSimonian and Laird [13] in which only an adjustment of the variance of the overall effect estimator is provided.

In our approach we use the model of unobserved heterogeneity as the starting point of all further analysis. We do this since the meta-likelihood for unobserved heterogeneity can't be increased any further for a given data set: it is the largest likelihood possible and provides what is called the *nonparametric maximum meta-likelihood* (NPMML) and the corresponding estimator is called the *nonparametric maximum meta-*

For internal use only

likelihood estimator (NPMMLE). To avoid a potential misunderstanding we point out that in the likelihood (5) m is treated as an *unknown* parameter. The NPMML is the maximum meta-likelihood for all possible values of $\lambda_1, ..., \lambda_m, p_1, ..., p_m$, *and* m. The meta-likelihood is *bounded* over the set of all discrete probability distributions on the real line, and, consequently, the NPMMLE exists. This is in contrast to other normal likelihoods where restrictions need to be placed to attain boundedness of the likelihood (an example is the mixture of normals with a common unknown variance parameter, where the likelihood increases beyond every bound when the number of components m is increasing). Here, these problems do not exist, and an estimate $\hat{m}$ of m exists which associates with the NPMMLE. Technically, using the EM algorithm this estimate is found in a conditional fashion, fixing the m to values 1, 2, 3, …, and then finding estimates with the EM algorithm for each value of m. Increasing m to (m + 1) is terminated when there is zero change in the associated likelihoods. Note that this is in contrast to other cases where a likelihood increase is continuing for all increases of m. Then NPMMLE (and with this $\hat{m}$) can also be computed with one of the existing global search algorithms like the vertex-exchange method [11].

3.4 Including Covariates to Explain Heterogeneity

Having identified considerable *unobserved* heterogeneity the question arises whether any *observed* variables can be associated with this *latent* form of heterogeneity. To put it in other words, one knows that there is heterogeneity, but it is yet unclear what it stands for. Having observed further covariates, $x_1, x_2, ..., x_p$, say, one can formulate a *regression model* to include these into the meta-likelihood

$$\prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{ - \tfrac{1}{2} (\hat{\lambda}_i-\lambda_i)^2/\sigma_i^2\}, \quad (8)$$

where now $\lambda_i = \beta_i x_{1i} + \beta_i x_{2i} + ... + \beta_i x_{pi} = \mathbf{x}_i^T \boldsymbol{\beta}$ is provided by the regression model. Finding the maximum likelihood estimate according to the meta-likelihood leads to the weighted regression estimator (which is provided for compactness in vector notation):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\, \mathbf{X}^T\mathbf{W}\mathbf{Y},$$

where

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ & & \dots. & & \\ 0 & 0 & \dots & 0 & w_k \end{pmatrix}$$

contain on the diagonal the inverse study variances $w_i = 1/\sigma_i^2$, $\mathbf{Y} = (\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_k)^T$ and $\mathbf{X}$ is the design matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ & & \dots. & \\ x_{k1} & x_{k2} & \dots & x_{kp} \end{pmatrix}$$

containing the study data of the p predictors in the k studies. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is easily available as $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$.

This powerful tool is readily available by means of any statistical package which can do weighted regression. Here, the package MINITAB [14] was used.

3.5 Model Evaluation

Various models are considered and need to be evaluated in terms of which model provides the most adequate explanation of the data. For the analysis provided here, the *Bayesian Information Criteria* was used throughout. It is defined as BIC = 2 meta-log-likelihood – N log(k), where *meta-log-likelihood* stands for the natural logarithm of the meta-likelihood of the model under consideration, N stands for the number of parameters involved in the model, and log(k) is the natural logarithm of the number of studies. The idea behind this criterion is that if one adds more parameters into the model the likelihood increases. It should therefore be penalized for the number of parameters involved in the model to balance the increase in fit with the increased model complexity. The BIC-criterion has turned out to be a valid instrument for discriminating models. For the homogeneity model, N = 1, since only the mean parameter is involved in the model. For the heterogeneity model, there are N = 2m – 1 parameters as mentioned above and for the regression model N = p, the number of covariates in the model. Having included all *relevant* covariates into the model one can expect that the log-likelihood for the regression model becomes close to the log-likelihood for the heterogeneity model, meaning that *most of the residual heterogeneity* has been explained. If this is *not* the case, it can be expected that some additional covariate (yet unknown) needs to be found for explaining the residual heterogeneity.

3.6 Publication Bias

To avoid drawing unbiased conclusions from a meta-analysis it is important that all relevant primary studies need to be identified on a given subject. It has been long accepted that research with statistically significant results is potentially more likely to be submitted, published or published more rapidly than work with null or non-significant results, leading to incorrect, usually effect-overestimating conclusions. This problem is known as *publication bias.* Methods are available for the diagnosis of publication bias including graphical methods such as the *funnel plot* [9] and statistical methods such as the rank correlation test [15], Rosenthal's 'file drawer' method [16], the more recent 'trim and fill' method [17], or regression techniques. A detailed discussion of these techniques can be found in [4, Ch. 7]. The basic idea of most of the techniques is based on the assumption that if there is *no* publication bias, then the effect measure should be unrelated to the sample size. If the sample size of the study is not available the surrogate 1/SE is used since it is known that the standard error is inversely related to the sample size. Though all of the above mentioned methods have their moments, in the approach here, we focus on regression methods since they allow the unifying treatment of the subject. We follow the ideas suggested in Macaskill et al. [18] in which the effect measure $\hat{\lambda}_i$ is regressed on $w_i = 1/\sigma_i^2$ using weights w_i. *If there is no publication bias, then the regression to the inverse vari-*

For internal use only

ance should show no effect. The benefit of this approach is that it can be simultaneously included in the previously mentioned regression approach for the covariates.

3.7 Study Quality

Various methodological issues can influence the quality of a study including study design (cohort, unmatched or matched case-control study, cross-sectional), case and control selection, cohort group selection, case and exposure assessment, and the kind of statistical analysis (parametric modelling or non-parametric estimation, logistic regression modelling or Mantel-Haenszel analysis). At best, none of these factors should be associated with the effect measure of interest. It has been suggested [4] to combine these individual markers into a quality score QS, and incorporate these scores as weights into the analysis. On the other hand, it is often remarked critically that these new weights might incorporate new, subjective choices into the meta-analysis, since different researchers dealing with the same body of evidence might come up with very different weighing schemes. In fact, Greenland [19] has indicated that quality assessment is the most insidious form of bias in the conduct of meta-analysis. It seems to exist a general agreement that a quality assessment of the primary studies should be carried out, possibly using a scale, checklist or individual components, though there is controversy how this should be incorporated at the analysis stage [20]. It is our opinion that the analysis should incorporate effects due to study quality, if agreement about study quality indicators can be reached. An example for such an agreement would be the fact if in a study an effect was adjusted for potential confounders or not; here most of the epidemiologists and statisticians would agree that it is very important to adjust for potential confounders.

It is therefore suggested [4, Ch. 8] to use a regression model in which the quality score is related to the effect measure. Here, we take up this straightforward idea, but extend it to allow other covariates in the model as well.

The idea can be formalized by regressing the estimate of the effect measure onto the quality score ($\hat{\lambda}_i$ on QS_i):

$$\hat{\lambda}_i = \alpha_0 + \alpha_1 QS_i + \beta^T \mathbf{x}_i + \varepsilon_i, \quad \text{for } i=1,\ldots,k \qquad (9)$$

where ε_i is a normal error with mean 0 and variance σ_i^2, and $\mathbf{x}_i$ the vector of covariates already included into the modelling. Estimates can again be found using weighted regression.

It might be argued that the construction of the quality score is a subjective instrument itself, especially if it is not well-accepted in the area of research at hand. Then, instead of using the quality score in (9), one simply uses the original marker variables describing the quality of a study. This results in a model with two kinds of covariate vectors:

$$\hat{\lambda}_i = \alpha^T \mathbf{s}_i + \beta^T \mathbf{x}_i + \varepsilon_i, \text{ for } i=1,\ldots,k \qquad (10)$$

where in contrast to (9) $\mathbf{s}_i$ is the vector of study quality covariates for study i.

4. Reanalysis of the Meta-Analysis by Sillero-Arenas et al. [1]

We now come back to the MA of SA and apply the ideas of section 3. The homogeneity model shows considerable difference to the heterogeneity model (see Table 4). Various models for heterogeneity have been estimated: models with m = 2, m = 3, and m = 4 subpopulations, the latter (m = 4) corresponding to the NPMMLE. The BIC-value suggests a model with m = 3 components (see Table 4). Therefore, it can be expected that covariates are to be found to explain the residual heterogeneity. We consider study type (case_control) and the number of covariates that has been adjusted for in each study (Number-of-Covariates). When we include the covariates one at a time, none of them are significant, though *Case-Control* and *Number-of-Covariates* are borderline (Table 3a). When we include these two simultaneously the latter becomes significant (see Table 3b). This also corresponds to a considerable amount of increase in the log-likelihood. Now, if the BIC-value of the regression model is compared with the BIC-model of the heterogeneity model, it is seen that the regression model provides the better BIC-value. Thus, it can be argued that most of the heterogeneity is explained. Note that the covariate *Number of covariates adjusted for* is transformed into a 0/1 covariate, indicating presence or absence of covariate adjustment in the original study. The basis for this transformation is provided in Figure 2 where there is almost no correlation visible for larger values of the covariate (right hand side of the Fig. 2).

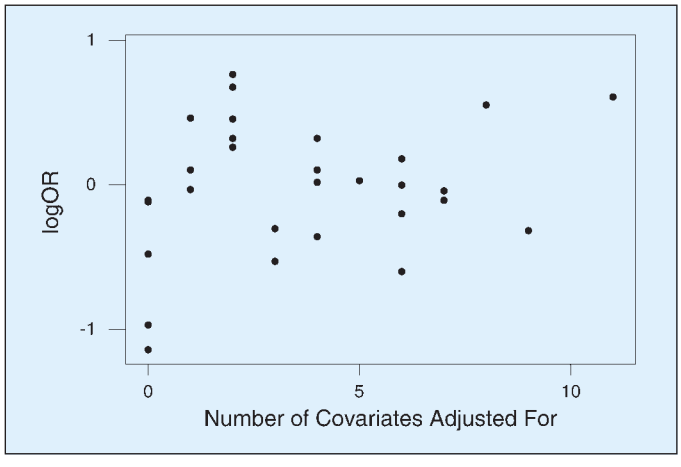

Fig. 2
Scatter plot of log OR vs. Number of covariates adjusted for

Internal use only

Table 3 Regression output forweighted regression of log-odds-ratio on study type (case-control) and covariate adjustment (number of covariates ajusted for)

a) ANALYSIS WITH STUDY_TYPE ONLY
The regression equation is logOR = 0.145 – 0.147 case_control

Predictor	Coef	StDev	T	P
Constant	0.14545	0.06601	2.20	0.034
case_cont	–0.14699	0.08664	–1.70	0.099

b) ANALYSIS WITH STUDY_TYPE AND NUMBER_OF_COVARIATES
The regression equation is logOR = – 0.229 + 0.356 number of covariates – 0.104 case_control
31 cases used 5 cases contain missing values or had zero weight

Predictor	Coef	StDev	T	P
Constant	–0.2288	0.1810	–1.26	0.217
number o	0.3564	00.1738	2.05	0.050
case_cont	–0.10384	0.09389	–1.11	0.278

c) ANALYSIS WITH STUDY_TYPE, NUMBER_OF_COVARIATES, AND WEIGHT (as replacement for sample size)
The regression equation is logOR = – 0,212 + 0,377 number of covariates – 0,120 case_control –0,000187 weight
31 cases used 5 cases contain missing values or had zero weight

Predictor	Coef	StDev	T	P
Constant	–0,2122	0,1890	–1,12	0,271
number o	0,3772	0,1849	2,04	0,051
case_cont	–0,1198	0,1043	–1,15	0,260
weight	–0,0001866	0,0004914	–0,38	0,707

Table 4 Measures to evaluate the various models: good models should have large BIC-value [BIC = 2 meta-log-likelihood – # parameters log (#studies)]

Model	Meta-log-likelihood	Number of Parameters	BIC
Homogeneity	–34.4630	1	–72.5095
mixture 2-components	–21.9740	3	–54.6986
mixture 3-components	–17.1960	5	–52.3096
mixture 4-components	–16.1373	7	–57.3592
Covariates	–19.2900	3	–49.3306

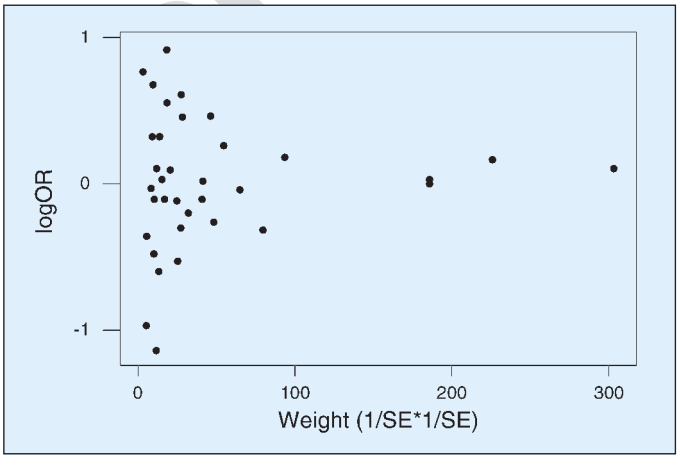

Fig. 3
Scatter plot (for diagnosis of publication bias) of log OR vs. weight

4.1 Estimated Adjusted Relative Risk

The estimated relative risk adjusted for study type and confounding variables can be found using the equation

log OR = – 0.229 + 0.356 number of covariates – 0.104 case_control

which leads to a log-odds ratio of 0.128 with 95% C.I. of (0.002-0.255) when *number of covariates* takes on the value 1 and *case_control* the value 0. This corresponds to an OR of 1.137 with 95% CI of (1.002-1.291).

4.2 Publication Bias

Neither the funnel plot (provided in Fig. 3) nor the weighted regression (weights equal to the inverse variance) of the log-odds ratio onto the inverse variance (weight) provide any evidence for presence of a publication bias. See Table 3c for the regression output.

Study quality was not further investigated, since the covariates describing it, namely *study type* and *adjustment for covariates,* have already been included into the model , and have proved to provide an effect. Therefore, study quality has inherently been taken into the modelling.

5. Discussion

In summary, we refocus on our approach. It is assumed that an effect measure is available which is normally distributed with *known* study-specific variances. This assumption is a mere working assumption and could be replaced by something else such as a Binomial or Poisson distribution for the measure of interest. We are using the normal model to retain the simplicity of presentation. In addition, the normal model provides often reasonable approximations for effect measures like risk differences or risk ratios, if those are appropriately transformed. To identify heterogeneity it is required to do a mixture analysis in the first place. If the

For personal use only

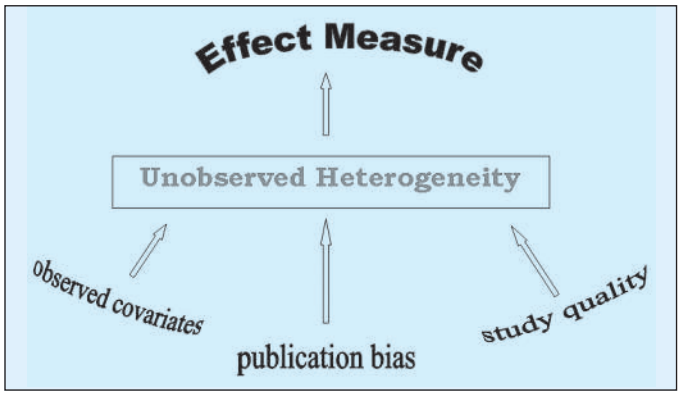

Fig. 4
Graphical summary of the general framework

meta-likelihood for models of homogeneity and heterogeneity agree, no further analysis is necessary, since there *is no* heterogeneity to be explained. In this case, one may proceed to the pooled analysis. Note that this implies as well that there is no need to check for publication bias nor study quality. In most cases, however, forms of heterogeneity might be found. These can be linked to study covariates describing study characteristics, study quality or publication bias. Effect estimates for these covariates can be simply found using weighted regression and significance of individual covariates using Wald statistics will lead to the final model. The idea is expressed in Figure 4 in a compact way. Finally, having identified an appropriate model one can compare the associated likelihood with the likelihood for the unobserved heterogeneity. If both likelihoods are close, then most of the heterogeneity has been accounted for. On the other hand, if there is still considerable disagreement between both likelihoods, the meta-analysis is still frail for explaining residual heterogeneity, potentially by means of a missing covariate, correlated studies, or other causes of extra-heterogeneity. With regard to this aspect, the MA could still be considered incomplete.

Coming to the MA of HRT and breast cancer, it is argued here that the overall result of the meta-analysis by SA which provided an odds ratio of 1.062 with 95% confidence interval (1.014–1.112) need to be corrected using up-to-date methods. Indeed, if we reconsider the meta-analysis by SA and use in a secondary analysis the tools available today, we find an estimate of 1.137 on the relative risk scale indicating a more elevated risk for HRT than the one provided by SA. This finding occurs since the meta-analysis at hand experiences considerable *unobserved heterogeneity,* which has previously been ignored. We argue further in this contribution that this form of heterogeneity can be successfully linked to covariates observed in the study base provided by SA, namely the study type (cohort or case-control) and whether in the study the estimate of effect has been adjusted for potential confounders. It can be furthermore established that these covariates do indeed correlate with the size of the effect measure. Consequently, an odds ratio adjusted for study type and confounder treatment seems to be more appropriate, leading to the one given above. This effect estimate is more in the direction and closer to the one of 1.26 provided in the WHI-trial [2].

In general, it appears appropriate and useful to include study characteristics (as well as patient characteristics) into the meta-analysis. The analysis tools are readily available and easy to handle. The problem though might be that not all of the *interesting* and *important* covariates might be available. Therefore, it is supremely important to *incorporate unobserved heterogeneity* into the meta-likelihood, which can be used as an indicator for covariates not yet known and/or not yet included into the meta-analysis. In the meta-analysis at hand most of the unobserved heterogeneity could be explained and this can be taken as considerable empirical evidence for the validity of the effect estimates found with the modelling approach taken here.

Furthermore, it was investigated if these results could be prone to any publication bias effect. Using the graphical device of a funnel plot in combination with an appropriate regression analysis no evidence of a presence of a publication bias could be detected.

Acknowledgments
This research was initiated during a 3-day workshop on Meta-Analysis which took place in July 2002 at the College of Public Health, University of the Philippines, Manila (Philippines) where the author was the principal lecturer. The workshop was excellently organized by Dr. Jesus Sarol, Jr. and numerous co-workers of his. I would like to express my sincere thanks to Dr. Sarol and all the people making this event possible.

References

1. Sillero-Arenas M, Delgado-Rodriguez M, Rodigues-Canteras R, Bueno-Cavanillas A, Galvez-Vargas R. Menopausal Hormone Replacement Thearapy and Breast Cancer: A Meta-Analysis. Obstet Gynecol 1992; 79: 286-94.
2. Rossouw JE et al. (Writing Group for the Women's Health Initiative Investigators). Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the women's health initiative randomized controlled trial. JAMA 2002; 288: 321-33.
3. Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russel Sage Foundation; 1994.
4. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-Analysis in Medical Research. Chichester: Wiley; 2000.
5. Altman D, Chalmers I, editors. Systematic Reviews. London: BMJ Publishing Group; 1995.
6. Petitti DB. Meta-analysis, Decision Analysis and Cost-Effectiveness Analysis. Methods for Quantitative Synthesis in Medicine. Oxford: Oxford University Press; 1994.
7. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. London: Academic Press; 1985.
8. Glass GV, McGraw B, Smith ML. Meta-analysis in Social Research. Newbury Park, CA: Sage; 1981.
9. Light RJ, Pillemar DB. Summing Up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press; 1984
10. Stangl DK, Berry DA, editors. Meta-analysis in Medicine and Health Policy. New York: Marcel Dekker; 2000.
11. Böhning D. Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis. Disease Mapping. and Others. Boca Raton: Chapman & Hall/CRC; 2000.
12. McLachlan G, Krishnan T. The EM Algorithm and Extensions. New York: Wiley; 1997.
13. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7: 177-88.
14. Minitab Inc. Minitab 13.3. State College: Minitab Inc.; 2000.

Non-commercial use only

15. Begg CB, Mazumdar M. Operator characteristics of a rank correlation test for publication bias. Biometrics 1994; 50: 1088-101.
16. Rosenthal R. The file drawer problem and tolerance for null results. Psycol Bull 1979; 86: 638-41.
17. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000; 56: 455-63.
18. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Stat Med 2001; 20: 641-54.
19. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 1994; 140: 290-6.
20. Berard A, Bravo G. Combining studies using effect sizes and quality scores: application to bone loss in postmenopausal women. J Clin Epidemiol 1998; 51: 801-7.

Correspondence to:
Dankmar Böhning
Division for International Health
Institute for Social Medicine, Epidemiology,
and Health Economics
Charité Medical School Berlin
Free University Berlin/Humboldt University at Berlin
Fabeckstr. 60–62
14195 Berlin
Germany
E-mail: dankmar.boehning@charite.de

For internal use only